5011-669(E)
2-110E

Statistics Canada Statistique Canada

Ottawa

2

MULTIPLE FRAME SAMPLING IN THE AES

- PROBLEMS IN APPLICATION

Barbara Armstrong
Institutional & Agriculture
Survey Methods Division

June 1980

TABLE OF CONTENTS

	PAGE
1. Introduction	1
2. Overlap Determination	1
3. Duplication in the List	3
References	6
APPENDIX 1: Unbiasedness of Rao's Estimators	7

1. INTRODUCTION

For the 1980 AES multiple frame sampling will be used in four of the six provinces covered by the survey. There are problems unique to multiple frame sampling which we have not encountered before in the AES. In this paper I will discuss two of these problems and how we will be dealing with them.

The first problem - that of overlap determination - is not new to us, nor is the proposed solution. We are familiar with both from the 1979 multiple frame test. However, I feel it advisable to document the rationale for our treatment of this problem for future reference.

The second problem is not so much a problem of multiple frame sampling but a problem of sampling from a list. The problem of duplication of units in a list sampling frame is discussed at length in a paper by Frederick Vogel (see reference 2) and I have drawn heavily from this paper in my discussion of the problem and the proposed solution.

2. OVERLAP DETERMINATION

In the 1980 AES we have sampled from two independent frames--one a list frame and the other an area frame. It is essential that the independence of the two frames be maintained to ensure that resulting multiple frame estimates will be unbiased. This is why we must be careful not to allow data from the area sample to update information on the list frame in rotation years. Except for specified farms, the list frame is fixed as of the redesign year (1979). Thus, for most of the list frame there is no danger of updating from the previous year's area sample. It is only for specified farms that we must ensure that no data obtained from the area sample will affect the selection of specified farms for the next year's survey (unless the survey is to be completely re-designed in the next year). For example, a large farm discovered in the area sample cannot be specified in the next year's survey.

In addition to the independence of survey frames, there are two other basic requirements of a multiple frame survey. These are:

1) Every unit belonging to the population of interest must appear on at least one of the sampling frames.

2) For every sampled unit it must be possible to determine on which sampling frame(s) the unit appears.

Use of an area frame ensures the complete population coverage required by condition (1).

On the list frame, the sampling unit is the name of a farm operator while the reporting unit is all the land that is operated by this farm operator. With the area frame, the sampling unit is a segment or piece of land while the reporting unit is every farm that has land inside this segment. To meet condition (2) every segment farm in the area sample must be identified by the name of its operator. We then rely on the reported names of farm operators to determine the overlap between the two survey frames.

The important point is that overlap between the list and area frames is achieved by matching names of farm operators not by matching pieces of farm land. Any change of farm ownership (e.g. sale of a farm, farm takeover by another family member) indicates that the original farm which appears on the list frame no longer exists. A change of ownership is said to have occurred when the reported name of the farm operator differs from the operator name on the list frame. The only exception is when the farm operator is a hired manager (usually in the case of a corporate farm) since a change in a hired manager does not necessarily reflect a change in farm ownership.

Suppose that a son, John Jones, has taken over the operation of his father's farm. The father, Bob Jones, appears on the list frame. If this farm were selected in the area sample the name of the farm operator, John Jones, would not be found on the list frame and would, therefore, contribute to the area non-overlap portion of the multiple frame estimate. If, however, Bob Jones were selected in the list sample an interviewer would go to the farm of Bob Jones and discover that it was now operated by John Jones. In this case, the farm on the list frame (i.e. the farm operated by Bob Jones) is treated as if it no longer exists. It is considered to be out of business. This treatment of list farms which have changed operators is necessary to ensure that such farms (which could be selected from either sampling frame) are not allowed to contribute to both the list and area non-overlap portion of the multiple frame estimate.

The exact criteria used to define an out of business list farm is this. If the name of the farm operator reported on the questionnaire differs from the name sampled from the list frame and the names of all reported partners differ from the name sampled from the list frame and the operator is not a hired manager then the list farm is said to be out of business.

It is often the case that a farm run in partnership has only one of the partners names reported on the list frame. Therefore, as long as the operator name appearing on the list frame appears as either the operator reported on the front page of the questionnaire or as one of the partners reported on the second page of the questionnaire, the farm is still considered the same as the sampled list farm. Similarly, in the area sample, when determining the overlap

status of segment farms, every partner's name is matched to the list frame. A match on any partner's name will result in the farm being classified as overlap.

Another factor which could complicate determination of overlap is leasing arrangements. However, as with partnerships consistent application of the out of business list farm status should result in the correct handling of such cases. If a farm operator has leased his land from the original operator who appears on the list frame then the situation is effectively the same as the case of a change in farm operator. If he falls in a sample segment, the operator leasing the land will not match with the list frame and will therefore contribute to the area non-overlap portion of the multiple frame estimate. If the original farm operator were sampled from the list frame, the interviewer would determine that the sampled farm was now run by a different operator (who is not simply a manager hired by the original operator). Thus this list farm would be declared out of business.

3. DUPLICATION IN THE LIST

The most common source of duplication on the list frame occurs when two or more list frame farms amalgamate into a single operation. This type of duplication becomes obvious when two or more sampled list farms have amalgamated into one. But the same problem exists if a sampled list farm amalgamates with a non-sampled list farm. In the first case the duplication is detected in the field. The fact that two or more sampled farms have amalgamated is noted on the questionnaires concerned and usually only one of the questionnaires is completed. In the second case (i.e. where a sampled farm has amalgamated with a non-sampled farm) a separate check is necessary to determine whether any of the partners reported for the sampled list farm appear separately on the list frame.

A procedure has been developed by J.N.K. Rao to handle this type of duplication on the list frame. To use this procedure we need to know for every selected unit the number of times it could have been selected (i.e. the number of times it appears on the list frame). Starting with the simple case where there is no stratification the ordinary estimator for the population total is

$$(1)\quad \hat{Y} = \frac{N}{n} \sum_{i=1}^{n} y_i$$ where N = number of units on the list frame
n = number of units sampled.

Suppose that there is duplication on the list frame. For every sampled unit i we have A_i, the number of times this unit appears on the list frame and hence could be sampled (A_i is equal to 1 if there is no duplication of the ith unit). Rao's estimate for Y is

$$(2)\quad \hat{Y} = \frac{N}{n} \sum_{i=1}^{n} \frac{y_i}{A_i} .$$

Proceeding to the case where there is stratification in the sample design, the usual estimator is

$$(3)\quad \hat{Y} = \sum_h \frac{N_h}{n_h} \sum_{i=1}^{n_h} y_{hi}$$

where N_h = number of units in stratum h

n_h = number of units sampled from stratum h.

Introducing duplication into the list, Rao's estimator is

$$(4)\quad \hat{Y} = \sum_h \frac{N_h}{n_h} \sum_{i=1}^{n_h} \frac{y_{hi}}{A_i}$$

where A_i is the number of times the ith unit in stratum h could have been selected from the list frame (not necessarily from the same stratum). See Appendix 1 for proof that (2) and (4) are unbiased estimators of the population total. The condition that must be met to ensure unbiasedness is that for each duplicated unit in the sample the value $\frac{y_{hi}}{A_i}$ is used for that unit in the tabulation of estimates.

The implication of this as far as AES processing goes is simply this. Every list sample farm must have a completed questionnaire even if two sample farms have amalgamated into one. This can be done by cross-referencing the completed list questionnaire to the amalgamated, incomplete questionnaire. This year a computer program will transfer data from the completed cross-referenced questionnaire to the incomplete questionnaire. (Note: Based on 1979 AES returns we know that in 1980 there are three cases where two list sample farms have amalgamated into one. In each of these cases only one questionnaire was sent to the field to be enumerated. The second questionnaire for each has been kept at Head Office. The AES ID will be entered on the second questionnaire as well as the cross-reference ID of the first questionnaire. These questionnaires can then be batched for data capture.)

Every list sample farm for which it has been determined that the farm appears on the list frame more than once must have its raising factor adjusted. If, for example, three list farms amalgamate into one and two of the three farms were selected in the list sample, then the raising factors for both of the sampled farms would be divided by three. (This is equivalent to dividing the farm's individual data values by three.)

If time permits these raising factor adjustments will be made to the raising factor file prior to running the raising factor match.

Otherwise preliminary estimates could be produced using the original raising factor file with revised estimates then produced once the raising factors have been adjusted. In 1979 such raising factor adjustments had a very minor impact on the level of the estimates. Hopefully this will continue to be the case in 1980.

Perhaps it is worth noting here why these problems will not arise in Ontario and Quebec since, strictly speaking, the AES design in these provinces is still a multiple frame type survey with an area sample being supplemented by a list sample of farms called specified farms. A list of large farms is formed (i.e. a list frame) in each of these provinces and is then enumerated completely in the AES. Because of this complete enumeration of specified (list) farms there is no problem in identifying overlap between the area sample and list frame. If a specified farm falls in a sample segment, this "overlap" will be determined by the interviewer during enumeration in the field. Such a farm contributes to the specified farm portion of the population estimate and is not included in the segment sample.

As for duplication of units on the list frame this, again, will be apparent during field enumeration. If two specified farms have amalgamated into one then only one questionnaire will be completed (for the whole amalgamated operation). Raising factors remain equal to 1.0 because specified farms belong to a complete enumeration stratum with probability of selection always equal to one.

REFERENCES

1. Rao, J.N.K., "Some Non-Response Sampling Theory When the Frame Contains an Unknown Amount of Duplication", JASA, March 1968 (87-90).

2. Vogel, F.A., "Surveys with Overlapping Frames--Problems in Application", presented at the 135th Annual Meeting of the ASA, August 25-28, 1975.

APPENDIX 1: Unbiasedness of Rao's Estimators

From (2) we have Rao's simple random sampling estimator with duplication on the list frame.

$$(2)\quad \hat{Y} = \frac{N}{n} \sum_{i=1}^{n} \frac{y_i}{A_i}$$

To show this is unbiased we can re-write (2) as

$$\hat{Y} = \frac{N}{n} \sum_{i=1}^{N} \frac{y_i}{A_i} t_i$$

where t_i is an indicator variable such that

$t_i = 0$ if the ith unit is not selected in the sample
$= 1$ if the ith unit is selected in the sample

We then know the expected value of t_i is

$$E(t_i) = \frac{n}{N} .$$

Therefore the expected value of $\hat{Y}$ is

$$E(\hat{Y}) = \frac{N}{n} \sum_{i=1}^{N} \frac{y_i}{A_i} E(t_i)$$

$$= \sum_{i=1}^{N} \frac{y_i}{A_i}$$

$$= \sum_{i=1}^{M} \left(\frac{y_i A_i}{A_i}\right)$$

$$= \sum_{i=1}^{M} y_i \quad \text{where } M = \text{number of unique units on the list frame.}$$

This estimator is unbiased as long as data for the duplicated unit is included in the tabulation of estimates every time it is selected in the sample.

Recall from (4), Rao's stratified estimator with duplication on the list frame is

$$(4)\quad \hat{Y} = \sum_h \frac{N_h}{n_h} \sum_{i=1}^{n_h} \frac{y_{hi}}{A_i} .$$

To show its unbiasedness we re-write (4) as

$$\hat{Y} = \sum_h \frac{N_h}{n_h} \sum_{i=1}^{N_h} \frac{y_{hi}}{A_i} \cdot t_{hi}$$

where t_{hi} is the indicator variable such that

t_{hi} = 0 if the ith unit in stratum h is not selected in the sample.

= 1 if the ith unit in stratum h is selected in the sample.

The expected value of t_{hi} is

$$E(t_{hi}) = \frac{n_h}{N_h} .$$

Then the expected value of Rao's estimator is

$$E(\hat{Y}) = \sum_h \frac{N_h}{n_h} \sum_{i=1}^{N_h} \frac{y_{hi}}{A_i} E(t_{hi})$$

$$= \sum_h \sum_{i=1}^{N_h} \frac{y_{hi}}{A_i} .$$

Now let A_{hi} be the number of times unit i in stratum h appears on the list frame in stratum h. Then

$$A_i = \sum_h A_{hi} .$$

We can then re-write the above equation as

$$E(\hat{Y}) = \sum_{h} \sum_{i=1}^{M_h} \frac{y_{hi}}{A_i} A_{hi} \quad \text{where } M_h = \text{number of unique units in stratum h}$$

$$= \sum_{i=1}^{M} y_i \, .$$

2-110F

2

L'ÉCHANTILLONNAGE À BASES MULTIPLES DANS L'EDA

PROBLÈMES PRATIQUES

Barbara Armstrong
Division des méthodes
d'enquête - institutions et agriculture

Juin 1980

TABLE DES MATIÈRES

	PAGE
1. Introduction	1
2. Détermination du chevauchement	1
3. Dédoublement dans la liste	3
Bibliographie	6
ANNEXE 1: L'absence de biais des estimations de Rao	7

1. INTRODUCTION

On utilisera lors de l'enquête descriptive sur l'agriculture de 1980 un échantillon à bases multiples dans quatre des six provinces enquêtées. L'échantillonnage à bases multiples pose des problèmes uniques que nous n'avons pas encore rencontrés dans l'EDA. Nous examinerons ici deux de ces problèmes et la façon dont nous allons les traiter.

Le premier problème, celui de la détermination du chevauchement, n'est pas nouveau pour nous, pas plus que la solution proposée. Le test à bases multiples de 1979 les renfermait déjà. Toutefois, j'estime qu'il est nécessaire de documenter notre solution de ce problème pour référence future.

Le deuxième problème n'est pas tellement un problème d'échantillonnage à bases multiples mais plutôt un problème d'échantillonnage à partir d'une liste. Le problème du dédoublement des unités d'une base de sondage par listes est abordé en détail dans un document de Grederick Vogel (voir référence bibliographique 2), et je m'en suis inspiré largement pour la préparation du présent document lors de l'examen du problème et de la solution proposée.

2. DETERMINATION DU CHEVAUCHEMENT

Dans l'EDA de 1980, nous avons tiré un échantillon de deux bases de sondage indépendantes, l'une qui est une base de sondage par listes, et l'autre une base de sondage aréolaire. Il est essentiel de garantir l'indépendance des deux bases afin d'éliminer la possibilité d'un biais des estimations ainsi obtenues. Pour cette raison, nous allons faire très attention à ce que les données provenant de l'échantillon aréolaire ne mettent pas à jour les renseignements de la base de sondage par listes lors des années de rotation. À l'exception de certaines fermes spécifiées, la base de sondage par listes reste la même que celle de l'année de changement de conception (1979). Le danger d'une mise à jour pour la plus grande partie de la base de sondage par listes à partir de l'échantillon aréolaire de l'année précédente n'existe pas. Ce n'est que pour les fermes spécifiées que nous devons garantir qu'aucune donnée obtenue à partir de l'échantillon aréolaire n'affectera la sélection des fermes spécifiées pour l'enquête de l'année suivante (à moins que l'enquête ne soit complètement reconçue l'année suivante). Ainsi, une grande ferme découverte dans l'échantillon aréolaire ne peut être spécifiée dans l'enquête de l'année suivante.

En plus de l'indépendance des bases de sondage, les enquêtes à bases multiples retiennent deux autres conditions fondamentales:

1) Chaque unité appartenant à la population étudiée doit figurer dans au moins une des bases de sondage.

2) Pour chaque unité échantillonnée, il doit être possible de déterminer dans quelle base (ou bases) elle figure.

L'utilisation d'une base aréolaire garantit une observation exhaustive de la population exigée par la condition (1).

Dans le cas de la base de sondage par liste, l'unité d'échantillonnage est le nom d'un exploitant agricole tandis que l'unité déclarante est l'ensemble du terrain qu'il exploite. Dans le cas de la base aérolaire, l'unité d'échantillonnage est un segment ou une parcelle de terrain, tandis que l'unité déclarante est chaque ferme qui possède un terrain à l'intérieur de ce segment. Afin de répondre à la condition (2), chaque ferme de segment de l'échantillon aréolaire doit être identifiée par le nom de son exploitant. Nous utilisons les noms déclarés des exploitants agricoles afin de déterminer le chevauchement entre les deux bases de sondage.

Ce dont il faut se souvenir ici, c'est que le chevauchement entre les deux bases de sondage s'obtient par une comparaison des noms des exploitants agricoles, et non pas par un rapprochement des parcelles de terre agricole. Tout changement de propriété (par exemple, la vente d'une ferme, une ferme reprise par un autre membre de la famille) signifie que la ferme originale qui figure dans la base de sondage par listes n'existe plus. On dit qu'il y a changement de propriété lorsque le nom déclaré de l'exploitant agricole diffère de celui figurant dans la base de sondage par listes. La seule exception est lorsque l'exploitant agricole est un gérant (habituellement dans le cas d'une ferme consituée en corporation), puisqu'un changement de gérant ne traduit pas nécessairement un changement de propriété pour la ferme en question.

Supposons qu'un fils, Jean Dupont, a pris en charge l'exploitation de la ferme de son père. Le père, Pierre Dupont, figure dans la base de sondage par listes. Si cette ferme avait été sélectionnée à partir de l'échantillon aréolaire, le nom de l'exploitant agricole, Jean Dupont, ne figurerait pas dans la base de sondage par listes et par conséquent, se rattacherait à la partie de non-chevauchement aréolaire de l'estimation établie à partir de la base de sondage multiple. Si toutefois Pierre Dupont avait été sélectionné à partir de la base de sondage par listes, un enquêteur irait voir à la ferme de Pierre Dupont et constaterait que celle-ci est maintenant exploitée par Jean Dupont. Dans ce cas, la ferme figurant dans la base de sondage par listes (c'est-à-dire la ferme exploitée par Pierre Dupont) est traitée comme si elle n'existait plus. On la considère comme hors exploitation. Ce traitement des fermes de liste qui ont changé d'exploitant est nécessaire afin de garantir que ces fermes (qui peuvent être choisies à partir de n'importe quelle base de sondage) ne contribuent pas simultannément à la partie de non-chevauchement des bases de sondage aérolaire et par listes lors de l'estimation de la base de sondage multiple.

Voici les critères exacts utilisés pour définir une ferme de liste hors exploitation. Si le nom de l'exploitant agricole déclaré sur le questionnaire diffère du nom échantillonné à partir de la base de

de sondage par listes et si les noms de tous les partenaires déclarés diffèrent du nom échantillonné à partir de la base de sondage par listes et si l'exploitant n'est pas un gérant, alors on dira que la ferme de liste est hors exploitation.

Il arrive souvent qu'un seul des noms des partenaires d'une ferme exploitée en association figure dans la base de sondage par listes. Par conséquent, tant que le nom de l'exploitant figurant dans la base de sondage par listes apparaît soit comme l'exploitant déclaré sur la première page du questionnaire, soit comme un des partenaires énumérés à la deuxième page du questionnaire, on considère que la ferme est la même que celle de la liste d'échantillonnage. De même, dans le cas de l'échantillon aréolaire, on apparie le nom de chaque partenaire à celui de la base de sondage par listes lorsqu'on détermine le chevauchement éventuel des fermes de segment. S'il y a correspondance du nom du partenaire, la ferme sera classée comme chevauchement.

Une autre cause de complication dans la détermination des chevauchements sont les dispositions de location. Toutefois, comme dans le cas des associations, une utilisation cohérente des fermes de liste hors exploitation devrait régler le problème. Si un exploitant agricole a loué sa terre auprès de l'exploitant original qui figure dans la base de sondage par listes, alors la situation est effectivement la même que dans le cas d'un changement d'exploitant agricole. S'il fait partie d'un segment d'échantillon, l'exploitant qui loue la terre ne correspondra pas à celui de la base de sondage par listes, et contribuera par conséquent à la partie de non-chevauchement aréolaire de l'estimation à bases multiples. Si l'exploitant agricole original a été sélectionné à partir de la base de sondage par listes, l'enquêteur déterminera que la ferme échantillonnée est maintenant gérée par un autre exploitant (qui n'est pas simplement un gérant embauché par l'exploitant original). Cette ferme de liste sera alors considérée comme hors exploitation.

3. DEDOUBLEMENT DANS LA LISTE

La cause la plus courante d'un dédoublement dans la base de sondage par listes se produit lorsque deux ou plusieurs fermes de la base de sondage par listes fusionnent en une seule exploitation. Ce type de dédoublement est évident lorsque deux ou plusieurs fermes de listes échantillonnées ont fusionné en une seule. Mais le problème demeure si une ferme de liste échantillonnée fusionne avec une ferme de liste non échantillonnée. Dans le premier cas, le dédoublement est détecté sur le terrain. Le fait que deux ou plusieurs fermes échantillonnées ont fusionné est indiqué sur les questionnaires, et habituellement un seul de ces derniers est rempli. Dans le deuxième cas (c.-à.d. lorsqu'une ferme échantillonnée a fusionné avec une ferme non échantillonnée), il faut procéder à une vérification séparée afin de déterminer si l'un des partenaires déclarés pour la ferme de liste échantillonnée figure séparément dans la base de sondage par listes.

J.N.K. Rao a mis un point un procédé pour régler ce problème dans la base de sondage par listes. Pour l'utiliser, nous devons connaître pour chaque unité sélectionnée le nombre de fois qu'elle pourrait être sélectionnée (c.-à-d. le nombre de fois qu'elle figure dans la base de sondage par listes). En partant du cas simple, lorsqu'il n'y a pas de stratification, l'estimateur ordinaire du total de la population est

$$(1)\quad \hat{Y} = \frac{N}{n} \sum_{i=1}^{n} y_i$$

où N = nombre d'unités dans la base de sondage par listes
n = nombre d'unités échantillonnées.

Supposons que la base de sondage par listes contienne des dédoublements. Pour chaque unité échantillonné i, nous avons A_i, qui est le nombre de fois que cette unité figure dans la base de sondage par listes, et qui par conséquent pourrait être échantillonnée (A_i est égal à 1 s'il n'y a pas de dédoublement de la i-ième unité). L'estimateur de Rao pour Y est

$$(2)\quad \hat{Y} = \frac{N}{n} \sum_{i=1}^{n} \frac{y_i}{A_i}$$

Lorsque l'échantillon est stratifié, l'estimateur habituel est:

$$(3)\quad \hat{Y} = \sum_{h} \frac{N_h}{n_h} \sum_{i=1}^{n_h} y_{hi}$$

où N_h = nombre d'unités dans la strate h
n_h = nombre d'unités échantillonnés à partir de la strate h.

En introduisant le dédoublement dans la liste, l'estimateur de Rao est:

$$(4)\quad \hat{Y} = \sum_{h} \frac{N_h}{n_h} \sum_{i=1}^{n_h} \frac{y_{hi}}{A_i}$$

où A_i est le nombre de fois que la i-ième unité de la strate h pourrait être choisie à partir de la base de sondage par listes (pas nécessairement à partir de la même strate). On trouvera à l'annexe 1 la preuve que (2) et (4) sont des estimateurs sans biais du total de la population. La condition qu'il faut satisfaire pour garantir l'absence de biais est que pour chaque unité dédoublée dans l'échantillon, la valeur $\frac{y_{hi}}{A_i}$ est utilisée pour cette unité lors de la totalisation des estimations.

En ce qui concerne le traitement de l'EDA, les implications sont simplement ceci. Chaque ferme d'échantillon de listes doit avoir un questionnaire rempli même si deux fermes échantillonnées ont fusionné en une seule. On peut y arriver par un recoupement du questionnaire de listes rempli avec le questionnaire fusionné (incomplet). Cette année, un programme informatique va transférer les données provenant du questionnaire rempli recoupé au questionnaire incomplet. (Remarque: sur la base des déclarations de l'EDA de 1979, nous savons qu'en 1980 il y a eu trois cas où deux fermes d'échantillon de listes ont fusionné en une seule. Dans chacun de ces cas, un questionnaire seulement a été envoyé au bureau régional pour le dénombrement. Le deuxième questionnaire pour chacune des fermes a été conservé au bureau central. L'identification de l'EDA sera inscrit sur le second questionnaire, de même que l'identification de recoupement du premier questionnaire. Ces questionnaires peuvent être ensuite mis en lots pour la saisie des données).

Il faut corriger le facteur d'expansion de chaque ferme d'échantillon de listes dont on a déterminé qu'elle figure dans la base de sondage par listes plus d'une fois. Si, par exemple, trois fermes de listes fusionnent en une seule, et si deux des trois fermes ont été sélectionnées dans la base de sondage par listes, alors les facteurs d'expansion pour chacune des deux fermes échantillonnées seront divisés par trois. (Ceci équivaut à diviser les valeurs des données individuelles de la ferme par trois).

Si le temps le permet, ces corrections des facteurs d'expansion seront apportées au fichier des facteurs d'expansion avant de procéder au rapprochement de ces derniers. Sinon, il serait possible de produire des estimations provisoires à partir du fichier des facteurs d'expansion original, et des estimations révisées lorsque les facteurs d'expansion auront été ajustés. En 1979, les corrections de ces facteurs d'expansion ont eu un impact minime sur le niveau des estimations. On espère que ce sera également le cas en 1980.

Il convient peut-être de noter pourquoi ces problèmes ne se sont pas posés en Ontario et au Québec, puisqu'à vrai dire l'EDA dans ces provinces est toujours une enquête du type à bases multiples, où un échantillon aréolaire est complété par un échantillon de listes des fermes appelées fermes spécifiées. On dresse une liste des grandes fermes (c.-à-d. une base de sondage par listes) dans chacune de ces provinces, qui font l'objet d'un dénombrement exhaustif lors de l'EDA. Pour cette raison, il n'y a pas de problème lors de l'identification des chevauchements entre l'échantillon aréolaire et la base de sondage par listes. Si une ferme spécifiées tombe dans un segment d'échantillon, ce "chevauchement" sera déterminé par l'enquêteur lors du dénombrement sur le terrain. Une telle ferme va contribuer à la portion des fermes spécifiées de l'estimation de la population, et ne sera pas prise en compte dans l'échantillon par segment.

En ce qui concerne le dédoublement des unités dans la base de sondage par listes, ce problème, là encore, va apparaître lors du dénombrement sur le terrain. Si deux fermes spécifiées ont fusionné en une seule, alors on remplira un seul questionnaire (pour l'ensemble de l'entreprise fusionnée). Les facteurs d'expansion demeurent égaux à 1.0, car les fermes appartiennent à une strate d'énumération complète, pour laquelle la probabilité de sélection est toujours égale à un.

BIBLIOGRAPHIE

1. Rao, J.N.K., "Some Non-Response Sampling Theory When the Frame Contains an Unknown Amount of Duplication", JASA, mars 1968 (87-90).

2. Vogel, F.A., "Surveys with Overlapping Frames--Problems in Application", présenté à la 135e réunion annuelle de l'ASA, Août 25-28, 1975.

ANNEXE 1: L'absence de biais des estimateurs de Rao

À partir de (2), nous obtenons l'estimateur d'échantillonnage aléatoire simple de Rao avec dédoublement dans la base de sondage par listes.

$$(2)\quad \hat{Y} = \frac{N}{n} \sum_{i=1}^{n} \frac{y_i}{A_i}$$

Pour prouver l'absence de biais, nous pouvons récrire (2) sous la forme:

$$\hat{Y} = \frac{N}{n} \sum_{i=1}^{N} \frac{y_i}{A_i} t_i$$

où t_i est une variable indicatrice telle que

$t_i = 0$ si la i-ième unité n'est pas sélectionnée dans l'échantillon

$= 1$ si la i-ième unité est sélectionné dans l'échantillon.

Nous savons alors que l'espérance mathématique de t_i est:

$$E(t_i) = \frac{N}{n}$$

Par conséquent, l'espérance mathématique de $\hat{Y}$ est

$$E(\hat{Y}) = \frac{N}{n} \sum_{i=1}^{N} \frac{y_i}{A_i} E(t_i)$$

$$= \sum_{i=1}^{N} \frac{y_i}{A_i}$$

$$= \sum_{i=1}^{M} \left(\frac{y_i A_i}{A_i}\right)$$

$$= \sum_{i=1}^{M} y_i \qquad \text{où M = nombre d'unités uniques de la base de sondage par listes.}$$

Cet estimateur est sans biais tant que les données de l'unité dedoublée sont incluses dans la totalisation des estimations chaque fois que cette unité est sélectionnée dans l'échantillon.

La formule (4) nous avait donné l'estimateur stratifié de Rao avec dédoublement pour la base de sondage par listes:

$$(4)\quad \hat{Y} = \sum_h \frac{N_h}{n_h} \sum_{i=1}^{n_h} \frac{y_{hi}}{A_i}$$

Pour prouver l'absence de biais, nous récrivons (4) sous la forme:

$$\hat{Y} = \sum_h \frac{N_h}{n_h} \sum_{i=1}^{N_h} \frac{y_{hi}}{A_i} \cdot t_{hi}$$

où t_{hi} est une variable telle que:

$t_{hi} = 0$ si la i-ième unité de la strate h n'est pas sélectionnée dans l'échantillon.

$= 1$ si la i-ième unité de la strate h est sélectionnée dans l'échantillon.

L'espérance mathématique de t_{hi} est:

$$E(t_{hi}) = \frac{N_h}{N_h}$$

L'espérance mathématique de l'estimateur de Rao est alors:

$$E(\hat{Y}) = \sum_h \frac{N_h}{n_h} \sum_{i=1}^{N_h} \frac{y_{hi}}{A_i} E(t_{hi})$$

$$= \sum_h \sum_{i=1}^{N_h} \frac{y_{hi}}{A_i}$$

Posons maintenant A_{hi} comme le nombre de fois que l'unité i de la strate h figure dans la base de sondage par listes dans la strate h. Alors

$$A_i = \sum_h A_{hi}$$

Nous pouvons alors récrire l'équation ci-dessus sous la forme:

$$E(\hat{Y}) = \sum_h \sum_{i=1}^{M_h} \frac{y_{hi}}{A_i} A_{hi}$$ où M_h est le nombre d'unités uniques dans la strate h

$$= \sum_{i=1}^{M} y_i \quad .$$

POOR

QUALITY

ORIGINAL